# Analysis of Spectral Hamiltonian Boundary Value Methods (SHBVMs) for the numerical solution of ODE problems [*]

Pierluigi Amodio[†] Luigi Brugnano[‡] Felice Iavernaro[*]

**Abstract**

Recently, the numerical solution of stiffly/highly-oscillatory Hamiltonian problems has been attacked by using Hamiltonian Boundary Value Methods (HBVMs) as spectral methods in time. While a theoretical analysis of this spectral approach has been only partially addressed, there is enough numerical evidence that it turns out to be very effective even when applied to a wider range of problems. Here we fill this gap by providing a thorough convergence analysis of the methods and confirm the theoretical results with the aid of a few numerical tests.

**Keywords:** Spectral methods, Legendre polynomials, Hamiltonian Boundary Value Methods, HBVMs, SHBVMs.

**MSC:** 65L05, 65P10.

## 1 Introduction

In recent years, the efficient numerical solution of Hamiltonian problems has been tackled via the definition of the energy-conserving Runge-Kutta methods named Hamiltonian Boundary Value Methods (HBVMs) [10, 11, 12, 13]. Such methods have been developed along several directions (see, e.g., [19, 4, 14]), including Hamiltonian BVPs [1] and Hamiltonian PDEs (see, e.g., [6]): we also refer to the monograph [7] and to the recent review paper [8] for further details.

More recently, HBVMs have been used as spectral methods in time for solving highly-oscillatory Hamiltonian problems [18], as well as stiffly-oscillatory Hamiltonian problems [9] emerging from the space semi-discretization of Hamiltonian PDEs. Their spectral implementation is justified by the fact that this family of methods performs a projection of the vector field onto a finite dimensional subspace via a least squares approach based on the use of Legendre orthogonal polynomials [13]. This spectral approach, supported by a very efficient nonlinear iteration technique to handle the large nonlinear systems needed to advance the solution in time (see [12], [7, Chapter 4] and [8]), proved to be very effective. However, a thorough convergence analysis of HBVMs, used as spectral methods, was still lacking. In fact, when using large stepsizes, as is required by the spectral strategy, the notion of classical order of a method is not sufficient to explain the correct asymptotic behaviour of the solutions, so that a different analysis is needed, which is the main goal of the present paper.

[*] Please, cite this paper as:
P. Amodio, L. Brugnano, F. Iavernaro. *Numer. Algorithms* (2019) https://doi.org/10.1007/s11075-019-00733-7

[†] Dipartimento di Matematica, Università di Bari, Italy. {pierluigi.amodio,felice.iavernaro}@uniba.it

[‡] Dipartimento di Matematica e Informatica "U. Dini", Università di Firenze, Italy. luigi.brugnano@unifi.it

Moreover, the theoretical achievements will be numerically confirmed by applying the methods to a number of ODE-IVPs.

It is worth mentioning that early references where numerical methods were derived by projecting the vector field onto a finite dimensional subspace are, e.g., [27, 28, 3, 2] (a related reference is [36]). A similar technique, popular for solving oscillatory problems, is that of *exponential/trigonometrically fitted* methods and, more in general, *functionally fitted* methods [24, 33, 37, 31, 23, 35, 26, 22, 29, 30, 32].

With these premises, the structure of the paper is as follows: in Section 2 we analyze the use of spectral methods in time; in Section 3 we discuss the efficient implementation of the fully discrete method; in Section 4 we provide numerical evidence of the effectiveness of such an approach, confirming the theoretical achievements. At last, a few conclusions are reported in Section 5.

## 2 Spectral approximation in time

This section contains the main theoretical results regarding the spectral methods that we shall use for the numerical solution of ODE-IVPs which, without loss of generality, will be assumed in the form[1]

$$\dot{y}(t) = f(y(t)), \qquad y(0) = y_0 \in \mathbb{R}^m. \tag{1}$$

Hereafter, $f$ is assumed to be suitably smooth (in particular, we shall assume $f(z)$ to be analytic in a closed complex ball centered at $y_0$). We consider the solution of problem (1) on the interval $[0, h]$, where $h$ stands for the time-step to be used by a one-step numerical method. The same arguments will be then repeated for the subsequent integration steps. According to [13], we consider the expansion of the right-hand side of (1) along the shifted and scaled Legendre polynomial orthonormal basis $\{P_j\}_{j\ge0}$,

$$P_j \in \Pi_j, \qquad \int_0^1 P_i(x)P_j(x)\mathrm{d}x = \delta_{ij}, \qquad i, j = 0, 1, \ldots,$$

with $\Pi_j$ the set of polynomials of degree $j$ and $\delta_{ij}$ the Kronecker delta. One then obtains:

$$\dot{y}(ch) = f(y(ch)) \equiv \sum_{j\ge0} P_j(c)\gamma_j(y), \qquad c \in [0, 1], \tag{2}$$

with the Fourier coefficients $\gamma_j(y)$ given by

$$\gamma_j(y) = \int_0^1 P_j(\tau)f(y(\tau h))\mathrm{d}\tau, \quad j = 0, 1, \ldots. \tag{3}$$

We recall that:

$$\|P_j\| := \max_{x\in[0,1]} |P_j(x)| = \sqrt{2j+1}, \qquad \int_0^1 P_j(x)q(x)\mathrm{d}x = 0, \quad \forall q \in \Pi_{j-1}. \tag{4}$$

Let us now study the properties of the coefficients $\gamma_j(y)$ defined at (3). To begin with, we report the following results.

[1] In fact, if problem (1) is non autonomous, $t$ can be included in the state vector.

**Lemma 1** *Let $g : [0, h] \to V$, with $V$ a vector space, admit a Taylor expansion at 0. Then*

$$\int_0^1 P_j(c)g(ch)\mathrm{d}c = O(h^j).$$

Proof   See [13, Lemma 1]. □

**Corollary 1** *The Fourier coefficients defined in (3) satisfy: $\gamma_j(y) = O(h^j)$.*

We now want to derive an estimate which generalizes the result of Corollary 1 to the case where the stepsize $h$ is not small. For this purpose, hereafter we assume that the solution $y(t)$ of (1) admits a complex analytic extension in a neighbourhood of 0. Moreover, we shall denote by $\mathcal{B}(0, r)$ the closed ball of center 0 and radius $r$ in the complex plane, and $\mathcal{C}(0, r)$ the corresponding circumference. The following results then hold true.

**Lemma 2** *Let $P_j$ be the $j$-th shifted and scaled Legendre polynomial and, for $\rho > 1$, let us define the function*

$$Q_j(\xi) = \int_0^1 \frac{P_j(c)}{\xi - c}\mathrm{d}c, \qquad \xi \in \mathcal{C}(0, \rho). \tag{5}$$

*Then,*

$$\|Q_j\|_\rho := \max_{\xi\in\mathcal{C}(0,\rho)} |Q_j(\xi)| \le \sqrt{\frac{2}{j+1}}\frac{1}{(\rho-1)\rho^j}. \tag{6}$$

Proof   One has, for $|\xi| = \rho > 1$, and taking into account (4):

$$\begin{aligned} Q_j(\xi) &= \int_0^1 \frac{P_j(c)}{\xi - c}\mathrm{d}c \;=\; \xi^{-1}\int_0^1 \frac{P_j(c)}{1-\xi^{-1}c}\mathrm{d}c \;=\; \xi^{-1}\int_0^1 P_j(c)\sum_{\ell\ge0}\xi^{-\ell}c^\ell\,\mathrm{d}c \\ &= \xi^{-1}\sum_{\ell\ge j}\xi^{-\ell}\int_0^1 P_j(c)c^\ell\,\mathrm{d}c \;=\; \xi^{-j-1}\sum_{\ell\ge0}\xi^{-\ell}\int_0^1 P_j(c)c^{\ell+j}\,\mathrm{d}c. \end{aligned}$$

Passing to norms, one has:

$$\begin{aligned} \left|\int_0^1 P_j(c)c^{\ell+j}\,\mathrm{d}c\right| &\le \|P_j\|\int_0^1 c^{\ell+j}\mathrm{d}c \;=\; \frac{\sqrt{2j+1}}{j+\ell+1} \;\le\; \sqrt{\frac{2}{j+1}}\,, \\ \left|\xi^{-j-1}\sum_{\ell\ge0}\xi^{-\ell}\right| &\le \rho^{-j-1}\sum_{\ell\ge0}\rho^{-\ell} \;=\; \left[(\rho-1)\rho^j\right]^{-1}\,, \end{aligned}$$

from which (6) follows. □

**Lemma 3** *Let $g(z)$ be analytic in the closed ball $\mathcal{B}(0, r^*)$ of the complex plane, for a given $r^* > 0$. Then, for all $0 < h \le h^* < r^*$,*

$$g_h(\xi) := g(\xi h) \tag{7}$$

*is analytic in $\mathcal{B}(0, \rho)$, with*

$$\rho \equiv \rho(h) := \frac{r^*}{h} \ge \frac{r^*}{h^*} =: \rho^* > 1. \tag{8}$$

We are now in the position of stating the following result.[2]

**Theorem 1** *Assume that the function*

$$g(z) := f(y(z)) \tag{9}$$

*and $h$ in (2)–(3) satisfy the hypotheses of Lemma 3. Then, there exists $\kappa = \kappa(h^*) > 0$, such that* [3]

$$|\gamma_j(y)| \le \frac{\kappa}{\sqrt{j+1}}\rho^{-j}. \tag{10}$$

Proof  By considering the function (7) corresponding to (9), and with reference to the function $Q_j(\xi)$ defined in (5), one has that the parameter $\rho$, as defined in (8), is greater than 1 and, moreover, (see (3))

$$\begin{aligned}\gamma_j(y) &= \int_0^1 P_j(c)f(y(ch))\mathrm{d}c \;\equiv\; \int_0^1 P_j(c)g_h(c)\mathrm{d}c \;=\; \int_0^1 P_j(c)\left[\frac{1}{2\pi\mathrm{i}}\int_{\mathcal{C}(0,\rho)}\frac{g_h(\xi)}{\xi-c}\mathrm{d}\xi\right]\mathrm{d}c\\ &= \frac{1}{2\pi\mathrm{i}}\int_{\mathcal{C}(0,\rho)} g_h(\xi)\left[\int_0^1\frac{P_j(c)}{\xi-c}\mathrm{d}c\right]\mathrm{d}\xi \;\equiv\; \frac{1}{2\pi\mathrm{i}}\int_{\mathcal{C}(0,\rho)} g_h(\xi)Q_j(\xi)\,\mathrm{d}\xi.\end{aligned}$$

Then, passing to norms (see (6)),

$$|\gamma_j(y)| \le \rho\|g_h\|_\rho\|Q_j\|_\rho.$$

Moreover, observing that (see (9), (7), and (8)):

$$\|g_h\|_\rho := \max_{\xi\in\mathcal{C}(0,\rho)}|g_h(\xi)| \le \max_{\xi\in\mathcal{B}(0,\rho)}|g_h(\xi)| \equiv \max_{z\in\mathcal{B}(0,r^*)}|g(z)| =: \|g\|,$$

and using (6), one has, (see (8)):

$$|\gamma_j(y)| \le \frac{\|g\|}{(1-\rho^{-1})}\sqrt{\frac{2}{j+1}}\rho^{-j} \le \frac{\|g\|}{(1-(\rho^*)^{-1})}\sqrt{\frac{2}{j+1}}\rho^{-j},$$

from which (10) eventually follows. □

**Remark 1** *It is worth mentioning that, in the bound (10), the dependence on $h$ only concerns the parameter $\rho > 1$, via the expression (8), from which one infers that $\rho \sim h^{-1}$, for all $0 < h \le h^* < r^*$. This, in turn, is consistent with the result of Corollary 1, when $h \to 0$.*

Let us now consider a polynomial approximation to (2),

$$\dot\sigma(ch) = \sum_{j=0}^{s-1} P_j(c)\gamma_j(\sigma), \qquad c\in[0,1], \tag{11}$$

[2] The used arguments are mainly adapted from [21].

[3] Hereafter, for sake of clarity, we shall denote by $|\cdot|$ any convenient vector or matrix norm.

where $\gamma_j(\sigma)$ is defined according to (3) by formally replacing $y$ by $\sigma$, i.e.,

$$\gamma_j(\sigma) = \int_0^1 P_j(\tau) f(\sigma(\tau h))\mathrm{d}\tau, \quad j = 0, 1, \ldots, s-1. \tag{12}$$

Integrating term by term (11), and imposing the initial condition in (1), provide us with the polynomial approximation of degree $s$:

$$\sigma(ch) = y_0 + h\sum_{j=0}^{s-1}\int_0^c P_j(x)\mathrm{d}x\, \gamma_j(\sigma), \qquad c \in [0,1]. \tag{13}$$

We now want to assess the extent to which $\sigma(ch)$ approximates $y(ch)$, for $c \in [0,1]$. When $h \to 0$, it is known that $y(h) - \sigma(h) = O(h^{2s+1})$ (see, e.g., [7, 8, 13]). Nevertheless, we here discuss the approximation of $\sigma$ to $y$, in the interval $[0,h]$, when $h$ is *finite* and only assuming that the result of Theorem 1 is valid. The following result then holds true.[4]

**Theorem 2** *Let $y$ be the solution of (1), $\sigma$ be defined according to (13), and assume that $f(\sigma(z))$ is analytic in $\mathcal{B}(0,r^*)$, for a given $r^* > 0$. Then, for all $0 < h \le h^* < r^*$, there exist $M, \bar{M} > 0$, $M = M(h^*)$, $\bar{M} = \bar{M}(h^*)$, and $\rho > 1$, $\rho \sim h^{-1}$, such that:*

- $|\sigma(ch) - y(ch)| \le chM\,\rho^{-s}, \quad c \in [0,1]\,;$
- $|\sigma(h) - y(h)| \;\le\; h\bar{M}\,\rho^{-2s}\,.$

Proof  Let $y(t,\xi,\eta)$ denote the solution of the problem

$$\dot{y} = f(y), \quad t \ge \xi, \qquad y(\xi) = \eta,$$

and $\Phi(t,\xi)$ be the solution of the associated variational problem,

$$\dot{\Phi}(t,\xi) = f'(y(t,\xi,\eta))\Phi(t,\xi), \quad t \ge \xi, \qquad \Phi(\xi,\xi) = I,$$

having set $f'$ the Jacobian of $f$. Without loss of generality, we shall assume that the function

$$\hat{g}(z) := \Phi(h,z), \tag{14}$$

is also analytic in $\mathcal{B}(0,r^*)$. We also recall the following well-known perturbation results:

$$\frac{\partial}{\partial\eta}y(t,\xi,\eta) = \Phi(t,\xi), \qquad \frac{\partial}{\partial\xi}y(t,\xi,\eta) = -\Phi(t,\xi)f(\eta).$$

Consequently, from (12) and (13), one has:

[4] The proof uses arguments similar to those of [13, Theorem 4].

$$\begin{aligned}
\sigma(ch)-y(ch) &= y(ch,ch,\sigma(ch))-y(ch,0,\sigma(0)) \;=\; \int_0^{ch}\frac{\mathrm{d}}{\mathrm{d}t}y(ch,t,\sigma(t))\,\mathrm{d}t\\
&= \int_0^{ch}\left[\frac{\partial}{\partial\xi}y(ch,\xi,\sigma(t))\Big|_{\xi=t}+\frac{\partial}{\partial\eta}y(ch,t,\eta)\Big|_{\eta=\sigma(t)}\dot\sigma(t)\right]\mathrm{d}t\\
&= -h\int_0^c\Phi(ch,\tau h)\left[f(\sigma(\tau h))-\dot\sigma(\tau h)\right]\mathrm{d}\tau\\
&= -h\int_0^c\Phi(ch,\tau h)\left[\sum_{j\ge0}P_j(\tau)\gamma_j(\sigma)-\sum_{j=0}^{s-1}P_j(\tau)\gamma_j(\sigma)\right]\mathrm{d}\tau\\
&= -h\sum_{j\ge s}\left[\int_0^c P_j(\tau)\Phi(ch,\tau h)\mathrm{d}\tau\right]\gamma_j(\sigma). \qquad (15)
\end{aligned}$$

From the result of Theorem 1 applied to $g(z):=f(\sigma(z))$, we know that there exist $\kappa=\kappa(h^*)$ and $\rho>1$, $\rho\sim h^{-1}$, such that, for the Fourier coefficients defined in (12),

$$|\gamma_j(\sigma)|\le\frac{\kappa}{\sqrt{j+1}}\rho^{-j}. \qquad (16)$$

Moreover, (see (4)) $\|P_j\|=\sqrt{2j+1}$ and, considering that, for all $h\in(0,h^*]$,

$$\max_{x_1,x_2\in[0,h]}|\Phi(x_1,x_2)| \;\le\; \max_{x_1,x_2\in[0,h^*]}|\Phi(x_1,x_2)| \;=:\; \nu \;\equiv\; \nu(h^*),$$

one has that

$$\left|\int_0^c P_j(\tau)\Phi(ch,\tau h)\mathrm{d}\tau\right|\le c\nu\sqrt{2j+1}.$$

Consequently, the first statement follows from (15) by setting, with reference to the parameter $\rho^*$ defined in (8),

$$M=\frac{\nu\kappa\sqrt2}{1-(\rho^*)^{-1}},$$

since, for all $c\in[0,1]$:

$$\begin{aligned}
|\sigma(ch)-y(ch)| &\le ch\nu\kappa\sum_{j\ge s}\sqrt{\frac{2j+1}{j+1}}\rho^{-j} \;\le\; ch\nu\kappa\sqrt2\sum_{j\ge s}\rho^{-j}\\
&= ch\nu\kappa\sqrt2\frac{\rho^{-s}}{1-\rho^{-1}} \;\le\; ch\frac{\nu\kappa\sqrt2}{1-(\rho^*)^{-1}}\rho^{-s} \;\equiv\; chM\rho^{-s}.
\end{aligned}$$

To prove the second statement (i.e., when $c=1$), we observe that the result of Theorem 1 holds true also for the function (14) involved in the Fourier coefficients

$$\int_0^1 P_j(\tau)\Phi(h,\tau h)\mathrm{d}\tau.$$

Consequently, there exist $\kappa_1 = \kappa_1(h^*) > 0$, such that

$$\left|\int_0^1 P_j(\tau)\Phi(h,\tau h)\mathrm{d}\tau\right| \le \frac{\kappa_1}{\sqrt{j+1}}\rho^{-j}. \tag{17}$$

The second statement then follows again from (15) by setting

$$\bar{M} = \frac{\kappa_1\kappa}{1-(\rho^*)^{-2}},$$

so that, by using same steps as above:

$$\begin{aligned} |\sigma(h) - y(h)| &\le h\kappa_1\kappa\sum_{j\ge s}\frac{\rho^{-2j}}{j+1} \le h\kappa_1\kappa\sum_{j\ge s}\rho^{-2j} \\ &= h\kappa_1\kappa\frac{\rho^{-2s}}{1-\rho^{-2}} \le h\frac{\kappa_1\kappa}{1-(\rho^*)^{-2}}\rho^{-2s} \equiv h\bar{M}\rho^{-2s}. \ \square \end{aligned}$$

Let us now introduce the use of a finite precision arithmetic, with machine precision $u$, for approximating (2). Then, the best we can do is to consider the polynomial approximation (11)–(12) [5]

$$\dot{y}(ch) \doteq \dot{\sigma}(ch) = \sum_{j=0}^{s-1} P_j(c)\gamma_j(\sigma), \qquad c \in [0,1], \tag{18}$$

such that

$$|\gamma_s(\sigma)| < tol \cdot \max_{j<s} |\gamma_j(\sigma)|, \qquad tol \sim u. \tag{19}$$

Integrating (18), and imposing that $\sigma(0) = y_0$, then brings back to (13). We observe that because of (16), (19) may be approximately recast as

$$\sqrt{\frac{1}{s+1}}\rho^{-s} < tol \sim u, \tag{20}$$

where $\rho \sim h^{-1}$. Consequently, choosing $s$ such that (19) (or (20)) is satisfied, we obtain that:

- the polynomial $\sigma(ch)$ defined by (18) and (13) provides a uniformly accurate approximation to $y(ch)$, in the whole interval $[0,h]$, within the possibility of the used finite precision arithmetic;

- $\sigma(h)$ is a *spectrally accurate* approximation to $y(h)$. Moreover, in light of the second point of the result of Theorem 2, one has that the criterion (19) can be conveniently relaxed. In fact, making the ansatz (see (16) and (17)) $\kappa = \kappa_1$, one has that

$$|\sigma(h) - y(h)| \lesssim \frac{h\kappa^2}{1-\rho^{-2}}\rho^{-2s} \approx \frac{h(s+1)}{1-\rho^{-2}}|\gamma_s(\sigma)|^2. \tag{21}$$

  Imposing the approximate upper bound to be smaller than the machine epsilon $u$, one then obtains:

$$|\gamma_s(\sigma)| \lesssim \sqrt{\frac{u(1-\rho^{-2})}{h(s+1)}} \propto u^{1/2}, \tag{22}$$

[5] Hereafter, $\doteq$ means "equal within the round-off error level of the used finite precision arithmetic".

which is generally much less restrictive than (19).[6] Alternatively, by considering that the use of relatively large time-steps $h$ is sought, one can use $tol \sim u^{1/2}$ in (19), that is,

$$|\gamma_s(\sigma)| < tol \cdot \max_{j<s} |\gamma_j(\sigma)|, \qquad tol \sim u^{1/2}, \tag{23}$$

In other words, (21) means that the method maintains the property of *super-convergence*, which is known to hold when $h \to 0$, also in the case where the time-step $h$ is relatively large.

**Remark 2** *In particular, we observe that (19) (or (20) or (22)) can be fulfilled by varying the value of $s$, and/or that of the stepsize $h$, by considering that, by virtue of (8),*

$$\rho(h_{new}) \approx \rho(h_{old})\frac{h_{old}}{h_{new}},$$

*$h_{old}$ and $h_{new}$ being the old and new stepsizes, respectively.*

*It is worth mentioning that the result of Theorem 2 can be also used to define a stepsize variation, within a generic error tolerance $tol$, thus defining a strategy for the simultaneous order/stepsize variation.*

We conclude this section mentioning that, to gain efficiency, the criterion (19) for the choice of $s$ in (18) can be more conveniently changed to

$$|\gamma_{s-1}(\sigma)| \le tol \cdot \max_{j\le s-1} |\gamma_j(\sigma)|, \qquad tol \sim \rho\cdot u. \tag{24}$$

Similarly, the less restrictive criterion (22) can be approximately modified as:

$$|\gamma_{s-1}(\sigma)| \lesssim \rho\sqrt{\frac{u}{hs}} \propto u^{1/2},$$

or, alternatively, one uses $tol \sim u^{1/2}$ in (24). As is clear, computing the norms of the coefficients $\gamma_j(\sigma)$ permits to derive estimates for the parameters $\kappa$ and $\rho$ in (16), as we shall see later in the numerical tests.

## 3 SHBVMs

The approximation procedure studied in the previous section does not yet provide a numerical method, in that the integrals defining $\gamma_j(\sigma)$, $j = 0,\dots,s-1$, in (12)–(13) need to be computed. For this purpose, one can approximate them to within machine precision through a Gauss-Legendre quadrature formula of order $2k$ (i.e., the interpolatory quadrature rule defined at the zeros of $P_k$) with $k$ large enough. In particular, following the criterion used in [18, 9], for the double precision IEEE[7] we choose

$$k = \max\{20, s+2\}. \tag{25}$$

After that, we define the approximation to $y(h)$ as

$$y_1 := \sigma(h) \equiv y_0 + h\gamma_0(\sigma). \tag{26}$$

[6] This latter criterion was that used in [18] and [9].

[7] In such a case, the machine precision is $u \approx 10^{-16}$.

In so doing, one eventually obtains a HBVM$(k,s)$ method, which we sketch below. Hereafter, we shall refer to such a method as to *spectral HBVM (in short, SHBVM)*, since its parameters $s$ and $k$, respectively defined in (19) (or (20) or (22)) and (25), are aimed at obtaining a numerical solution which is accurate within the round-off error level of the used finite precision arithmetic.

For sake of completeness, let us now briefly sketch what a HBVM$(k,s)$ method is. In general, to approximate the Fourier coefficient $\gamma_j(\sigma)$, and assuming for sake of simplicity that full machine accuracy is gained, we use the quadrature

$$\gamma_j(\sigma) \doteq \sum_{\ell=1}^k b_\ell P_j(c_\ell) f(\sigma(c_\ell h)) \;=:\; \hat\gamma_j, \qquad j=0,\dots,s-1, \tag{27}$$

where the polynomial $\sigma$ is that defined in (13) by formally replacing $\gamma_j(\sigma)$ with $\hat\gamma_j$, and $(c_i,b_i)$ are the abscissae and weights of the Gauss-Legendre quadrature of order $2k$ on the interval $[0,1]$.[8] Setting $Y_\ell=\sigma(c_\ell h)$, from (27) one then obtains the *stage equations*

$$Y_i \;=\; y_0+h\sum_{j=0}^{s-1}\int_0^{c_i}P_j(x)\mathrm{d}x\hat\gamma_j \;\equiv\; y_0+h\sum_{j=1}^k \underbrace{b_j\left[\sum_{\ell=0}^{s-1}\int_0^{c_i}P_\ell(x)\mathrm{d}xP_\ell(c_j)\right]}_{=:\,a_{ij}} f(Y_j), \quad i=1,\dots,k, \tag{28}$$

with the new approximation given by (see (26))

$$y_1 \;=\; y_0+h\hat\gamma_0 \;\equiv\; y_0+h\sum_{i=1}^k b_i f(Y_i). \tag{29}$$

Consequently, with reference to (28), setting

$$A=(a_{ij})\in\mathbb{R}^{k\times k}, \qquad \boldsymbol{b}=(b_i),\;\boldsymbol{c}=(c_i)\in\mathbb{R}^k, \tag{30}$$

one easily realizes that (28) and (29) define the $k$-stage Runge-Kutta method with Butcher tableau:

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^\top \end{array}\;.$$

From (28) one verifies that the Butcher matrix in (30) can be written as

$$A=\mathcal{I}_s\mathcal{P}_s^\top\Omega, \tag{31}$$

with

$$\mathcal{P}_s=\left(\begin{array}{ccc} P_0(c_1) & \dots & P_{s-1}(c_1)\\ \vdots & & \vdots\\ P_0(c_k) & \dots & P_{s-1}(c_k)\end{array}\right),\; \mathcal{I}_s=\left(\begin{array}{ccc} \int_0^{c_1}P_0(x)\mathrm{d}x & \dots & \int_0^{c_1}P_{s-1}(x)\mathrm{d}x\\ \vdots & & \vdots\\ \int_0^{c_k}P_0(x)\mathrm{d}x & \dots & \int_0^{c_k}P_{s-1}(x)\mathrm{d}x\end{array}\right)\in\mathbb{R}^{k\times s}, \tag{32}$$

and

$$\Omega=\left(\begin{array}{ccc} b_1 & & \\ & \ddots & \\ & & b_k\end{array}\right)\in\mathbb{R}^{k\times k}. \tag{33}$$

[8] I.e., $0<c_1<\dots<c_k<1$ are the zeros of $P_k$.

In fact, setting $\boldsymbol{e}_i \in \mathbb{R}^k$ the $i$-th unit vector, and taking into account (31)–(33), one has

$$\boldsymbol{e}_i^\top A \boldsymbol{e}_j = \boldsymbol{e}_i^\top \mathcal{I}_s \mathcal{P}_s^\top \Omega \boldsymbol{e}_j = \boldsymbol{e}_i^\top \mathcal{I}_s \left(\boldsymbol{e}_j^\top \mathcal{P}_s\right)^\top b_j = b_j \left[\sum_{\ell=0}^{s-1} \int_0^{c_i} P_\ell(x) \mathrm{d}x P_\ell(c_j)\right] \equiv a_{ij},$$

as defined in (28). From well-known properties of Legendre polynomials (see, e.g., [7, Appendix A]), one has that

$$\mathcal{I}_s = \mathcal{P}_{s+1} \hat{X}_s \equiv \begin{pmatrix} P_0(c_1) & \dots & P_s(c_1) \\ \vdots & & \vdots \\ P_0(c_k) & \dots & P_s(c_k) \end{pmatrix} \begin{pmatrix} \xi_0 & -\xi_1 & & \\ \xi_1 & 0 & \ddots & \\ & \ddots & \ddots & -\xi_{s-1} \\ & & \xi_{s-1} & 0 \\ & & & \xi_s \end{pmatrix}, \quad \xi_i = \left(2\sqrt{|4i^2-1|}\right)^{-1}, \tag{34}$$

from which one easily derives the following property relating the matrices (32)–(33) (see, e.g., [7, Lemma 3.6]):

$$\mathcal{P}_s^\top \Omega \mathcal{I}_s = \begin{pmatrix} \xi_0 & -\xi_1 & & \\ \xi_1 & 0 & \ddots & \\ & \ddots & \ddots & -\xi_{s-1} \\ & & \xi_{s-1} & 0 \end{pmatrix} =: X_s \in \mathbb{R}^{s \times s}. \tag{35}$$

**Remark 3** *From (32)–(34), one has that the Butcher matrix (31) can be rewritten as*

$$A = \mathcal{P}_{s+1} \hat{X}_s \mathcal{P}_s^\top \Omega. \tag{36}$$

*Considering that, when $k = s$, (see (35)) $\mathcal{P}_{s+1}\hat{X}_s = \mathcal{P}_s X_s$ and $\mathcal{P}_s^\top \Omega = \mathcal{P}_s^{-1}$, so that $A$ reduces to $\mathcal{P}_s X_s \mathcal{P}_s^{-1}$, we observe that (36) can be also regarded as a generalization of the $W$-transformation in [25, Section IV.5].*

At this point, we observe that the stage equations (28) can be cast in vector form, by taking into account (30)–(33), as

$$Y \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = \boldsymbol{e} \otimes y_0 + h \mathcal{I}_s P_s^\top \Omega \otimes I_m \cdot f(Y), \qquad \boldsymbol{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^k, \tag{37}$$

with an obvious meaning of $f(Y)$. On the other hand, the block vector of the coefficients in (27) turns out to be given by

$$\hat{\boldsymbol{\gamma}} \equiv \begin{pmatrix} \hat{\gamma}_0 \\ \vdots \\ \hat{\gamma}_{s-1} \end{pmatrix} = \mathcal{P}_s^\top \Omega \otimes I_m \cdot f(Y). \tag{38}$$

Consequently, from (37) one obtains

$$Y = \boldsymbol{e} \otimes y_0 + h \mathcal{I}_s \otimes I_m \cdot \hat{\boldsymbol{\gamma}},$$

and then, from (38), one eventually derives the equivalent discrete problem

$$F(\hat{\boldsymbol{\gamma}}) \,:=\, \hat{\boldsymbol{\gamma}} - \mathcal{P}_s^\top\Omega\otimes I_m\cdot f\left(\boldsymbol{e}\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\hat{\boldsymbol{\gamma}}\right) \,=\, \mathbf{0}, \tag{39}$$

which has (block) dimension $s$, independently of $k$ (compare with (37)). Once it has been solved, the new approximation is obtained (see (29)) as $y_1 = y_0 + h\hat{\gamma}_0$.

It is worth observing that the new discrete problem (39), having block dimension $s$ independently of $k$, allows us to use arbitrarily high-order quadratures (see (25)), without affecting that much the computational cost.

In order to solve (39), one could in principle use a fixed-point iteration,[9]

$$\hat{\boldsymbol{\gamma}}^{\ell+1} := \mathcal{P}_s^\top\Omega\otimes I_m\cdot f\left(\boldsymbol{e}\otimes y_0 + h\mathcal{I}_s\otimes I_m\cdot\hat{\boldsymbol{\gamma}}^\ell\right), \qquad \ell = 0,1,\dots,$$

which, though straightforward, usually implies restrictions on the choice of the stepsize $h$. For this reason, this approach is generally not useful when using the methods as spectral methods, where the use of relatively large stepsizes is sought. On the other hand, the use of the simplified Newton iteration for solving (39) reads, by virtue of (35),

$$\text{solve: } \left[I_s\otimes I_m - hX_s\otimes f'(y_0)\right]\boldsymbol{\delta}^\ell = -F(\hat{\boldsymbol{\gamma}}^\ell), \qquad \hat{\boldsymbol{\gamma}}^{\ell+1} := \hat{\boldsymbol{\gamma}}^\ell + \boldsymbol{\delta}^\ell, \qquad \ell = 0,1,\dots. \tag{40}$$

However, the coefficient matrix in (40) has a dimension $s$ times larger than that of the continuous problem (i.e., $m$) and, therefore, this can be an issue when large values of $s$ are to be used, as in the case of SHBVMs. Fortunately, this problem can be overcome by replacing the previous iteration (40) with a corresponding *blended iteration* [12, 7, 8] (see also [5]). In more details, once one has formally computed the $m\times m$ matrix

$$\Sigma = \left(I_m - h\rho_s f'(y_0)\right)^{-1}, \qquad \rho_s = \min_{\lambda\in\sigma(X_s)}|\lambda|, \tag{41}$$

where $\sigma(X_s)$ denotes, as is usual, the spectrum of matrix $X_s$, one iterates:

$$\boldsymbol{\eta}^\ell := F(\hat{\boldsymbol{\gamma}}^\ell), \quad \boldsymbol{\eta}_1^\ell := \rho_s X_s^{-1}\otimes I_m\boldsymbol{\eta}^\ell, \quad \hat{\boldsymbol{\gamma}}^{\ell+1} := \hat{\boldsymbol{\gamma}}^\ell + I_s\otimes\Sigma\left[\boldsymbol{\eta}_1^\ell + I_s\otimes\Sigma\left(\boldsymbol{\eta}^\ell - \boldsymbol{\eta}_1^\ell\right)\right], \qquad \ell = 0,1,\dots. \tag{42}$$

Consequently, one only needs to compute, at each time-step, the matrix $\Sigma\in\mathbb{R}^{m\times m}$ defined in (41),[10] having the same size as that of the continuous problem. Moreover, it is worth mentioning that for semi-linear problems with a leading linear part, the Jacobian of $f$ can be approximated with the (constant) linear part, so that $\Sigma$ is computed once for all [6, 18, 9].

**Remark 4** *It must be stressed that it is the availability of the very efficient blended iteration (41)–(42) which makes the practical use of HBVMs as spectral methods in time possible, since relatively large values of $s$ can be easily and effectively handled.*

*A thorough analysis of the blended iteration can be found in [15]. Contexts where it has been successfully implemented include stiff ODE-IVPs [16], linearly implicit DAEs up to index 3 [17] (see also the code* `BiMD` *in TestSet for IVP Solvers [40]), and canonical Hamiltonian systems (see the Matlab code* `HBVM`*, available at [39]), while its implementation in the solution of RKN methods may be found in [38].*

[9] Hereafter, the initial approximation $\hat{\boldsymbol{\gamma}}^0 = \mathbf{0}$ is conveniently used.

[10] I.e., factor $\Sigma^{-1}$.

# 4 Numerical tests

The aim of this section is twofold: firstly, to assess the theoretical analysis of SHBVMs made in Section 2; secondly, to compare such methods w.r.t. some well-known ones. All numerical tests, which concern different kinds of ODE problems, have been computed on a laptop with a 2.8GHz Intel-i7 quad-core processor and 16GB of memory, running Matlab 2017b. For the SHBVM method, the criterions (23) and (25) have been respectively used to determine its parameters $s$ and $\kappa$.

### The Kepler problem

We start considering the well-known Kepler problem (see, e.g., [7, Chapter 2.5]), which is Hamiltonian, with Hamiltonian function

$$H(q,p) = \frac{1}{2}\|p\|_2^2 - \|q\|_2^{-1}, \qquad q,p \in \mathbb{R}^2. \tag{43}$$

Consequently, we obtain the equations

$$\dot{q} = p, \qquad \dot{p} = -\|q\|_2^{-3} q, \tag{44}$$

which, when coupled with the initial conditions

$$q(0) = \begin{pmatrix} 1-\varepsilon, & 0 \end{pmatrix}^\top, \qquad p(0) = \begin{pmatrix} 0, & \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \end{pmatrix}^\top, \qquad \varepsilon \in [0,1), \tag{45}$$

provide a periodic orbit of period $T = 2\pi$ that, in the $q$-plane, is given by an ellipse of eccentricity $\varepsilon$. In particular, we choose the value $\varepsilon = 0.5$. The solution of this problem has two additional (functionally independent) invariants besides the Hamiltonian (43), i.e., the *angular momentum* and one of the nonzero components of the Lenz vector [7, page 64] (in particular, we select the second one):

$$M(q,p) = q_1p_2 - p_1q_2, \qquad L(q,p) = -p_1M(q,p) - q_2\|q\|_2^{-1}. \tag{46}$$

At first, we want to assess the result of Theorem 1. For this purpose, we apply the HBVM(20,16) method for one step starting from the initial condition (45), and using time-steps $h_i = 2\pi/(5\cdot 2^{i-1})$, $i = 1,\dots,5$. In Figure 1 is the plot (see (27)) of $|\hat{\gamma}_j|$, for $j = 0,1,\dots,15$, (solid line with circles), which, according to (16), should behave as $\kappa\rho^{-j}/\sqrt{j+1}$, due to the result of Theorem 1. A least squares approximation technique has been employed to estimate the two parameters $\kappa$ and $\rho$ appearing in the bound (16). These theoretical bounds are highlighted by the dashed line with asterisks in Figure 1: evidently, they well fit the computed values, except those which are close to the round-off error level. Moreover, according to the arguments in the proof of Theorem 1, one also expects that the estimate of $\kappa$ increases but is bounded, as $h \to 0$, whereas $\rho$ should be proportional to $h^{-1}$. This fact is confirmed by the results listed in Table 1.

Next, we compare the following methods for solving (44)–(45):

- the $s$-stage Gauss method (i.e., HBVM($s,s$)), $s = 1,2$, which is symplectic and of order $2s$. Consequently, it is expected to conserve the angular momentum $M(q,p)$ in (46), which is a quadratic invariant [34];
- the HBVM(6,$s$) method, $s = 1,2$, which, for the considered stepsizes is energy-conserving and of order $2s$;

Table 1: estimated values for the parameters $\kappa$ and $\rho$ for the Kepler problem (44)–(45), when using decreasing time-steps.

| $h$ | $\kappa$ | $\rho$ |
|---|---|---|
| $2\pi/5$ | 9.4 | 2.2 |
| $2\pi/10$ | 14.3 | 3.2 |
| $2\pi/20$ | 18.2 | 5.6 |
| $2\pi/40$ | 21.5 | 10.6 |
| $2\pi/80$ | 16.1 | 19.8 |

- the SHBVM method described above, where $s$ and $k$ are determined according to (23) and (25), respectively, with $tol \approx 10^{-8}$. This tolerance, in turn, should provide us with full accuracy, according to the result of Theorem 2, because of the super-convergence of the method, which is valid for any used step-size.[11]

It is worth mentioning that the execution times that we shall list for the Gauss, HBVM, and SHBVM methods are perfectly comparable, since the same Matlab code has been used for all such methods. This code, in turn, is a slight variant of the `hbvm` function available at the url [39].

In Tables 2–4, we list the obtained results when using a time-step $h = 2\pi/n$ over 100 periods. In more details, we list the maximum errors, measured at each period, in the invariants (43) and (46), $e_H, e_M, e_L$, respectively, the solution error, $e_y$, and the execution times (in sec). As it is expected, the symplectic methods conserve the angular momentum (since it is a quadratic invariant), whereas the energy-conserving HBVMs conserve the Hamiltonian function.[12] On the other hand, the SHBVM conserves all the invariants, and has a uniformly small solution error, by using very large stepsizes. Further, its execution time is the lowest one (less than 0.5 sec, when using $h = 2\pi/5$), thus confirming the effectiveness of the method.

### A Lotka-Volterra problem

We consider the following Poisson problem [20],

$$\dot{y} = B(y)\nabla H(y), \qquad B(y)^\top = -B(y), \tag{47}$$

with $y \in \mathbb{R}^3$ and, for arbitrary real constants $a, b, c, \nu, \mu$,

$$B(y) = \begin{pmatrix} 0 & c\,y_1y_2 & bc\,y_1y_3 \\ -c\,y_1y_2 & 0 & -y_2y_3 \\ -bc\,y_1y_3 & y_2y_3 & 0 \end{pmatrix}, \qquad H(y) = ab\,y_1 + y_2 - a\,y_3 + \nu\ln y_2 - \mu\ln y_3. \tag{48}$$

Moreover, assuming that $abc = -1$, there is a further invariant besides the Hamiltonian $H$, i.e., the *Casimir*

$$C(y) = ab\ln y_1 - b\ln y_2 + \ln y_3. \tag{49}$$

[11] As matter of fact, considering more stringent tolerances does not improve the accuracy of the computed numerical solution.

[12] In this case, the Gauss methods exhibit a super-convergence in the conservation of the Hamiltonian (3 times the usual order) and HBVMs do the same with the angular momentum. This is due to the fact that the error is measured only at the end of each period.

Table 2: numerical result for the $s$-stage Gauss method, $s = 1, 2$, used for solving the Kepler problem (44)–(45), $\varepsilon = 0.5$, with stepsize $h = 2\pi/n$.

| Gauss-1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | time | $e_H$ | rate | $e_M$ | $e_L$ | rate | $e_y$ | rate |
| 100 | 2.52 | 6.56e-03 | — | 5.88e-15 | 4.97e-01 | — | 3.04e 00 | — |
| 200 | 4.74 | 1.63e-03 | 2.0 | 1.04e-14 | 3.54e-01 | 0.5 | 2.39e 00 | 0.3 |
| 400 | 9.78 | 3.82e-04 | 2.1 | 2.09e-14 | 9.76e-02 | 1.9 | 1.61e 00 | 0.6 |
| 800 | 17.26 | 3.05e-05 | 3.6 | 7.66e-15 | 2.45e-02 | 2.0 | 7.49e-01 | 1.1 |
| 1600 | 33.74 | 6.07e-07 | 5.6 | 1.93e-14 | 6.11e-03 | 2.0 | 2.08e-01 | 1.8 |
| 3200 | 65.46 | 9.65e-09 | 6.0 | 3.04e-14 | 1.53e-03 | 2.0 | 5.25e-02 | 2.0 |
| Gauss-2 | | | | | | | | |
| $n$ | time | $e_H$ | rate | $e_M$ | $e_L$ | rate | $e_y$ | rate |
| 50 | 2.33 | 2.05e-06 | — | 3.44e-15 | 3.81e-02 | — | 3.17e-01 | — |
| 100 | 4.09 | 5.37e-10 | 11.9 | 5.77e-15 | 2.43e-03 | 4.0 | 2.09e-02 | 3.9 |
| 200 | 7.52 | 1.44e-13 | 11.9 | 7.55e-15 | 1.53e-04 | 4.0 | 1.32e-03 | 4.0 |
| 400 | 14.48 | 9.55e-15 | 3.9 | 9.99e-14 | 9.55e-06 | 4.0 | 8.29e-05 | 4.0 |
| 800 | 26.75 | 1.53e-14 | *** | 1.49e-14 | 5.97e-07 | 4.0 | 5.18e-06 | 4.0 |
| 1600 | 52.46 | 3.81e-14 | *** | 1.95e-14 | 3.73e-08 | 4.0 | 3.24e-07 | 4.0 |
| 3200 | 101.74 | 3.49e-14 | *** | 4.71e-14 | 2.33e-09 | 4.0 | 2.04e-08 | 4.0 |

Table 3: numerical result for the HBVM(6,$s$) method, $s = 1, 2$, used for solving the Kepler problem (44)–(45), $\varepsilon = 0.5$, with stepsize $h = 2\pi/n$.

| HBVM(6,1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | time | $e_H$ | $e_M$ | rate | $e_L$ | rate | $e_y$ | rate |
| 100 | 3.55 | 4.44e-16 | 9.09e-04 | — | 4.99e-01 | — | 2.94e 00 | — |
| 200 | 7.10 | 4.44e-16 | 2.12e-05 | 6.0 | 3.52e-01 | 1.9 | 9.68e-01 | 1.9 |
| 400 | 12.47 | 6.66e-16 | 3.39e-07 | 6.0 | 9.70e-02 | 2.0 | 2.58e-01 | 2.0 |
| 800 | 22.86 | 4.44e-16 | 5.29e-09 | 6.0 | 2.44e-02 | 2.0 | 6.46e-02 | 2.0 |
| 1600 | 45.46 | 4.44e-16 | 8.26e-11 | 6.0 | 6.10e-03 | 2.0 | 1.62e-02 | 2.0 |
| 3200 | 86.34 | 6.66e-16 | 1.30e-12 | 6.0 | 1.53e-03 | 2.0 | 4.04e-03 | 2.0 |
| HBVM(6,2) | | | | | | | | |
| $n$ | time | $e_H$ | $e_M$ | rate | $e_L$ | rate | $e_y$ | rate |
| 50 | 2.92 | 4.44e-16 | 1.09e-07 | — | 3.82e-02 | — | 4.64e-02 | — |
| 100 | 4.50 | 4.44e-16 | 2.72e-11 | 12.0 | 2.43e-03 | 4.0 | 2.94e-03 | 4.0 |
| 200 | 8.10 | 4.44e-16 | 5.88e-15 | 12.1 | 1.53e-04 | 4.0 | 1.84e-04 | 4.0 |
| 400 | 15.48 | 4.44e-16 | 3.89e-15 | *** | 9.55e-06 | 4.0 | 1.15e-05 | 4.0 |
| 800 | 28.42 | 4.44e-16 | 1.40e-14 | *** | 5.97e-07 | 4.0 | 7.20e-07 | 4.0 |
| 1600 | 52.29 | 6.66e-16 | 1.73e-14 | *** | 3.73e-08 | 4.0 | 4.50e-08 | 4.0 |
| 3200 | 107.41 | 6.66e-16 | 1.40e-14 | *** | 2.33e-09 | 4.0 | 2.81e-09 | 4.0 |

Table 4: numerical result for the SHBVM method used for solving the Kepler problem (44)–(45), $\varepsilon = 0.5$, with stepsize $h = 2\pi/n$.

| $n$ | $k$ | $s$ | time | $e_H$ | $e_M$ | $e_L$ | $e_y$ |
|---|---|---|---|---|---|---|---|
| 5 | 24 | 22 | 0.47 | 4.44e-16 | 2.01e-14 | 1.66e-14 | 8.00e-13 |
| 10 | 20 | 16 | 0.71 | 4.44e-16 | 6.22e-15 | 2.34e-14 | 6.13e-13 |
| 20 | 20 | 11 | 1.22 | 4.44e-16 | 6.66e-16 | 3.89e-15 | 3.87e-13 |
| 40 | 20 | 9 | 2.16 | 2.22e-16 | 1.89e-15 | 3.28e-15 | 5.75e-13 |

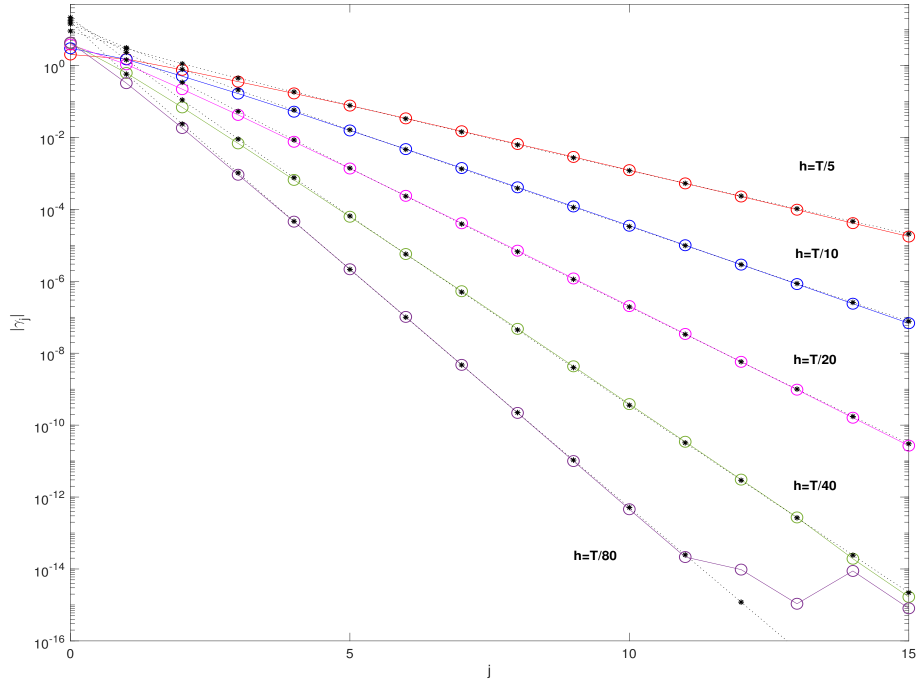

Figure 1: behavior of $|\hat{\gamma}_j|$ for decreasing values of the time-step $h$ for the Kepler problem (44)–(45) solved by the HBVM(20,16) method with decreasing time-steps. The line with circles are the computed norms, whereas those with the asterisks are the estimated ones. Observe that, for the smallest time-steps, the computed norms stagnate near the round-off error level.

The solution turns out to be periodic, with period $T \approx 2.878130103817$, when choosing

$$a = -2, \quad b = -1, \quad c = -0.5, \quad \nu = 1, \quad \mu = 2, \quad y(0) = (1,\, 1.9,\, 0.5)^\top. \tag{50}$$

For this problem, the HBVM($k, s$) method is no more energy-conserving, as well as the $s$-stage Gauss method. As matter of fact, both exhibit a drift in the invariants and a quadratic error growth in the numerical solution. The obtained results for the SHBVM method, with $tol \approx 10^{-8}$ in (23) for choosing $s$, $\kappa$ given by (25), and using a stepsize $h = T/n$, are listed in Table 5, where it is reported the maximum Hamiltonian error, $e_H$, the Casimir error, $e_C$, and the solution error $e_y$, measured at each period, over 100 periods. In such a case, all the invariants turn out to be numerically conserved, and the solution error is uniformly very small. Moreover, the SHBVM using the largest time-step (i.e., $h = T/5 \approx 0.57$) turns out to be the most efficient one. For comparison, in the table we also list the results obtained by using the Matlab solver `ode45` used with the default parameters, requiring 5600 integration steps and stepsizes approximately in the range $[2.2 \cdot 10^{-2}, 1.1 \cdot 10^{-1}]$, and the same solver used with parameters `AbsTol=1e-15`, `RelTol=1e-10`, now requiring 121760 integration steps, with stepsizes approximately in the range $[10^{-3}, 4.2 \cdot 10^{-3}]$.

Table 5: numerical result for the SHBVM method used for solving the Lotka-Volterra problem (47)–(50) with stepsize $h = T/n$. We also list the results obtained by using `ode45`, both with the default parameters, and with parameters `AbsTol=1e-15`, `RelTol=1e-10` (which we denote by `ode45*`).

| $n$ | $k$ | $s$ | time | $e_H$ | $e_C$ | $e_y$ |
|---|---|---|---|---|---|---|
| 5 | 20 | 16 | 0.88 | 8.26e-14 | 4.89e-14 | 4.24e-11 |
| 10 | 20 | 11 | 1.37 | 1.33e-14 | 1.33e-14 | 5.01e-11 |
| 15 | 20 | 9 | 1.84 | 3.11e-14 | 1.62e-14 | 4.92e-11 |
| `ode45` | | | 0.23 | 7.41e-01 | 7.27e-01 | 3.62e 00 |
| `ode45*` | | | 4.12 | 1.14e-08 | 8.71e-09 | 8.44e-07 |

Table 6: numerical result for the SHBVM method used for solving the stiff problem (51)–(52) with stepsize $h = 100/n$, and `ode15s` with the default parameters.

| $n$ | $k$ | $s$ | time | $e_y$ |
|---|---|---|---|---|
| 50 | 40 | 38 | 0.09 | 2.92e-11 |
| 75 | 32 | 30 | 0.12 | 1.53e-11 |
| 100 | 28 | 26 | 0.17 | 1.93e-12 |
| 125 | 25 | 23 | 0.21 | 6.28e-12 |
| 150 | 22 | 20 | 0.27 | 9.43e-12 |
| `ode15s` | | | 0.68 | 3.76e-04 |

**A stiff ODE-IVP**

At last, we consider a stiff ODE-IVP,

$$\dot{y}(t) = \begin{pmatrix} -9999 & 1 & 1 \\ 9900 & -100 & 1 \\ 98 & 98 & -2 \end{pmatrix} \big[y(t) - g(t)\big] + \dot{g}(t), \qquad y(0) = g(0), \tag{51}$$

with $g(t)$ a known function, having evidently solution $y(t) = g(t)$. We choose

$$g(t) = \begin{pmatrix} \cos 2\pi t, & \cos 4\pi t, & \cos 6\pi t \end{pmatrix}^\top, \tag{52}$$

and consider the SHBVM with $tol \approx 10^{-8}$ in (23) for choosing $s$ (as before, $\kappa$ is chosen according to (25)), so that full accuracy is expected in the numerical solution. The time-step used is $h = 100/n$ for $n$ steps. The measured errors in the last point (coinciding with the initial condition), are then reported in Table 6. For comparison, also the results obtained by the Matlab solver `ode15s`, using its default parameters, are listed in the table. This latter solver requires 6006 steps, with time-steps approximately in the range $[1.9 \cdot 10^{-3},\ 2 \cdot 10^{-2}]$.

# 5 Conclusions

In this paper we provide a thorough analysis of SHBVMs, namely HBVMs used as spectral methods in time, which further confirms their effectiveness. From the analysis, one obtains that the super-convergence of HBVMs is maintained also when using relatively large time-steps. SHBVMs become a practical method, due to the very efficient nonlinear *blended* iteration inherited from HBVMs. As a consequence, SHBVMs appear to be good candidates as *general ODE solvers.* This is indeed confirmed by a few numerical tests concerning a Hamiltonian problem, a Poisson (not Hamiltonian) problem, and a stiff ODE-IVP. The same tests show the numerical assessment of the theoretical achievements.

## Acknowledgements

The authors are very grateful to two unknown referees, for the careful reading of the manuscript, and for their precious comments, which allowed to formulate in a cleaner and more precise way the results presented in the paper.

# References

[1] P. Amodio, L. Brugnano, F. Iavernaro. Energy-conserving methods for Hamiltonian Boundary Value Problems and applications in astrodynamics. *Adv. Comput. Math.* **41** (2015) 881–905.

[2] P. Betsch, P. Steinmann. Inherently energy conserving time finite elements for classical mechanics. *J. Comp. Phys.* **160** (2000) 88–116.

[3] C.L. Bottasso. A new look at finite elements in time: a variational interpretation of Runge–Kutta methods. *Appl. Numer. Math.* **25** (1997) 355–368.

[4] L. Brugnano, M. Calvo, J.I. Montijano, L. Rández. Energy preserving methods for Poisson systems. *J. Comput. Appl. Math.* **236** (2012) 3890–3904.

[5] L. Brugnano, G. Frasca Caccia, F. Iavernaro. Efficient implementation of Gauss collocation and Hamiltonian Boundary Value Methods. *Numer. Algorithms* **65** (2014) 633–650.

[6] L. Brugnano, G. Frasca Caccia, F. Iavernaro. Line Integral Solution of Hamiltonian PDEs. *Mathematics* **7**(3) (2019) article n. 275. `https://doi.org/10.3390/math7030275`

[7] L. Brugnano, F. Iavernaro. *Line Integral Methods for Conservative Problems.* Chapman and Hall/CRC, Boca Raton, FL, 2016.

[8] L. Brugnano, F. Iavernaro. Line Integral Solution of Differential Problems. *Axioms* **7**(2) (2018) article n. 36. `http://dx.doi.org//10.3390/axioms7020036`

[9] L. Brugnano, F. Iavernaro, J.I. Montijano, L. Rández. Spectrally accurate space-time solution of Hamiltonian PDEs. (2018) `https://doi.org/10.1007/s11075-018-0586-z`

[10] L. Brugnano, F. Iavernaro, D. Trigiante. Hamiltonian BVMs (HBVMs): A family of "drift-free" methods for integrating polynomial Hamiltonian systems. *AIP Conf. Proc.* **1168** (2009) 715–718.

[11] L. Brugnano, F. Iavernaro, D. Trigiante. Hamiltonian Boundary Value Methods (Energy Preserving Discrete Line Integral Methods). *JNAIAM J. Numer. Anal. Ind. Appl. Math.* **5**, 1-2 (2010) 17–37.

[12] L. Brugnano, F. Iavernaro, D. Trigiante. A note on the efficient implementation of Hamiltonian BVMs. *J. Comput. Appl. Math.* **236** (2011) 375–383.

[13] L. Brugnano, F. Iavernaro, D. Trigiante. A simple framework for the derivation and analysis of effective one-step methods for ODEs. *Appl. Math. Comput.* **218** (2012) 8475–8485.

[14] L. Brugnano, F. Iavernaro, D. Trigiante. A two-step, fourth-order method with energy preserving properties. *Computer Phys. Commun.* **183** (2012) 1860–1868.

[15] L. Brugnano, C. Magherini. Blended Implementation of Block Implicit Methods for ODEs. *Appl. Numer. Math.* **42** (2002) 29–45.

[16] L. Brugnano, C. Magherini. The `BiM` Code for the Numerical Solution of ODEs. *J. Comput. Appl. Math.* **164–165** (2004) 145–158.

[17] L. Brugnano, C. Magherini, F. Mugnai. Blended Implicit Methods for the Numerical Solution of DAE Problems. *J. Comput. Appl. Math.* **189** (2006) 34–50.

[18] L. Brugnano, J.I. Montijano, L. Rández. On the effectiveness of spectral methods for the numerical solution of multi-frequency highly-oscillatory Hamiltonian problems. *Numer. Algorithms* (2018) `http://dx.doi.org/10.1007/s11075-018-0552-9`

[19] L. Brugnano, Y. Sun. Multiple invariants conserving Runge-Kutta type methods for Hamiltonian problems. *Numer. Algorithms* **65** (2014) 611–632.

[20] D. Cohen, E. Hairer. Linear energy-preserving integrators for Poisson systems. *BIT* **51** (2011) 91–101.

[21] P.J. Davis. *Interpolation & Approximations.* Dover, New York, 1975.

[22] J.M. Franco. Exponentially fitted symplectic integrators of RKN type for solving oscillatory problems. *Comput. Phys. Comm.* **177**, no. 6 (2007) 479–492.

[23] J.M. Franco. Runge-Kutta methods adapted to the numerical integration of oscillatory problems. *Appl. Numer. Math.* **50** (2004) 427–443.

[24] W. Gautschi. Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math.* **3** (1961) 381–397.

[25] E. Hairer, G. Wanner. *Solving Ordinary Differential Equations II, 2nd revised edition.* Springer, Heidelberg, 2002.

[26] N.S. Hoang, R.B. Sidje, N.H. Cong. On functionally-fitted Runge-Kutta methods. *BIT* **46** (2006) 861–874.

[27] B.L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comp.* **26** (1972) 415–426.

[28] B.L. Hulme. Discrete Galerkin and related one-step methods for ordinary differential equations. *Math. Comp.* **26** (1972) 881–891.

[29] J. Martin-Vaquero, J. Vigo-Aguiar. Exponential fitted Gauss, Radau and Lobatto methods of low order. *Numer. Algorithms* **48**, no. 4 (2008) 327–346.

[30] J.I. Montijano, M. Van Daele, M. Calvo. Functionally fitted explicit two step peer methods. *J. Sci. Comput.* **64**, no. 3 (2015) 938–958.

[31] Z. Kalogiratou, T.E. Simos. Construction of trigonometrically and exponentially fitted Runge-Kutta-Nyström methods for the numerical solution of the Schrödinger equation and related problem -a method of 8th algebraic order. *J. Math. Chem.* **31**, no. 2 (2002) 211–232.

[32] Y.-W. Li, X. Wu. Functionally fitted energy-preserving methods for solving oscillatory nonlinear Hamiltonian systems. *SIAM J. Numer. Anal.* **54**, no. 4 (2016) 2036–2059.

[33] B. Paternoster. Runge-Kutta (-Nyström) methods for ODEs with periodic solutions based on trigonometric polynomials. *Appl. Numer. Math.* **28** (1998) 401–412.

[34] J.M. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT* **28**, no. 4 (1988) 877–883.

[35] T.E. Simos. A trigonometrically-fitted method for long-time integration of orbital problems. *Math. Comput. Modelling* **40**, no. 11-12 (2004) 1263–1272.

[36] W. Tang, Y. Sun. Time finite element methods: a unified framework for numerical discretizations of ODEs. *Appl. Math. Comp.* **219** (2012) 2158–2179.

[37] G. Vanden Berghe, H. De Meyer, M. Van Daele, T. Van Hecke. Exponentially-fitted explicit Runge-Kutta methods. *Comput. Phys. Commun.* **123** (1999) 7–15.

[38] B. Wang, F. Meng, Y. Fang. Efficient implementation of RKN-type Fourier collocation methods for second-order differential equations. *Appl. Numer. Math.* **119** (2017) 164–178.

[39] `http://web.math.unifi.it/users/brugnano/LIMbook/`

[40] `https://archimede.dm.uniba.it/~testset/testsetivpsolvers/`